

Digital Data Management Plan of the Fermilab Short-Baseline Neutrino Program: SBND and ICARUS Detectors (DRAFT: v0.2 27 July 2022)

1. Introduction

The ICARUS detector is currently in operation while the SBND detector is to begin operations in 2023. As such, precise details of ICARUS data formats are available but unnecessary for the purpose of the Digital Data Management Plan. SBND data formats as well as specific publication procedures and policies can be expected to evolve prior to the start of data collection. This document outlines the basic elements of a plan for managing future digital data from the program and will evolve appropriately as the operational conditions change.

The Short-Baseline Near Detector (SBND) and ICARUS collaborations agree with the principles affirmed by the DOE Office of Science and the National Science Foundation related to the management of digital research data. This document describes general principles for managing the experimental digital data that satisfy the requirements expressed in the DOE Statement on Digital Data Management and following the Additional Requirements and Guidance for Digital Data Management as well as the NSF Policies on the Dissemination and Sharing of Research Results.

2. Data Management Resources and Facilities

SBND and ICARUS data management will be defined in accordance with an agreement between the Fermilab Scientific Computing Division and the SBND and ICARUS collaborations, as will be detailed in a Technical Scope of Work (TSW) document to be drafted and agreed upon before the start of data taking. Fermilab resources will provide a means to store, manage, access, and share the raw data and reconstruction data, as well as all of the research dependent analysis and calibration data.

3. SBND and ICARUS Data

While precise details of SBND data formats are still to be determined, the data for both ICARUS and SBND may be conceptually categorized into different “Tiers” based on the volume, their source, the required processing and selection criteria, and the expertise required to consume or reproduce the data. Substantial quantities of simulated data will also be generated as an important ingredient in the analysis of real data collected by the detectors. The policies outlined here apply to both real and simulated forms of data.

Each Tier is described in the subsections below. Tiers are listed in order of derivation: Raw Data, Reconstruction Data, Analysis Data, Published Data. The nominal policy on sharing and preservation of data in each Tier and how these data can be validated are given. Data sharing is considered to either be among members of the collaboration or between the collaboration and non-members. Preservation only considers copies of data shared within the collaboration. Requests for any expansion beyond the nominal policies described may be considered by the collaboration at any time on a case-by-case basis.

3.1 Raw Data

The Raw Data Tier includes all files produced directly from experiment devices (e.g., detector DAQ, environment monitors, beam monitors) and files holding the information used to configure these devices. The bulk of this Tier’s data will come from the SBND and ICARUS detectors and consist of digitized signals from TPC wires, optical detectors, and muon detectors in custom packed binary formats that require special software to be read. Also included in the Raw Data Tier is information about the beam and environment held in relational databases.

Raw Data is only shared among collaborators. The volume, infrastructure, and expertise required to produce and consume this data makes sharing outside the collaboration impractical. In

principle, all collaborators have access to this data but in practice only a few are expected to access a small portion of it. The primary consumer of this data is the official collaboration production processing.

All data in this Tier is archived to tape storage at Fermilab for the lifetime of the experiment and at least 5 years after data taking ceases as per the official Fermilab Data Management Plan. Additionally, the ICARUS raw data will be replicated to tape storage at CERN. Plans for permanent preservation will be made at the time when data taking ceases in order to utilize appropriate technological choices.

Validation of the Raw Data Tier is largely done by validating the proper operation of the devices that acquired it. This is done through detector commissioning and special-purpose studies as well as continual monitoring of the data acquisition by human shift operators.

3.2 Reconstruction Data

The Reconstruction Data Tier consists of files derived from Raw Data. It consists of intermediate results from processes such as noise reduction, signal extraction, imaging, pattern recognition, vertex and particle identification, as well as derived calibrations. The volume of data in this Tier is at least as large as Raw Data. Data in this Tier is shared following the same policies as the Raw Data Tier.

Data in the Reconstruction Data Tier is preserved to disk and tape at least until it is superseded by newer processing and is no longer actively utilized for measurements. Typically, long term preservation is not cost effective as Reconstruction Data can be reproduced by rerunning the software. Validation of this data is performed by comparisons between its similar derivations from Raw Data.

3.3 Analysis Data

An Analysis Data Tier consists of a down-sampling of the Reconstruction Data. Selection criteria are applied that reduce which quantities and triggers are kept. Expertise not generally available to the general public is needed to interpret the data in this tier. Data in this Tier is shared following the same policy as the Raw Data Tier.

Processes producing this data are relatively simple and validation is done by collaborators to assure the selection criteria perform as expected.

3.4 Published Data

Published Data consists of files directly used to produce the tables and figures used in published documents. The volume of data is relatively small, typically in formats that are readable with common tools (including ROOT), and hold quantities which may be properly interpreted by an individual with general understanding of the field. Files of Published Data can be made available at the time of publication along with the digital document, either through references given in the document or by request to the collaboration.

Published Data files will be preserved by the collaboration for the lifetime of the collaboration. Preservation policy of files shared through an external service will be determined by that service. Validation of these files will be done through the publication policy and procedures of the collaboration.

4. Information Protection

SBND and ICARUS data do not contain any personal identifiable or other confidential information, and therefore the information covered by this plan is not encumbered with any of the qualifiers listed in requirement four of the Statement on Digital Data Management. The exception being that the data is consider proprietary to the SBND and ICARUS Collaborations, and therefore the collaborations have a responsibility to assure sufficient expertise is employed in interpreting all but the Published Data. For this and practical reasons, access to non-published data is limited to

collaboration members and through specific agreements with other collaborations and scientists.

DRAFT